How to evaluate the degree of interdisciplinarity of an institution?

Lorenzo Cassi, Wilfriedo Mescheba, Elisabeth de Turckheim

Authors final version of the paper published in
Scientometrics 101 (3) DOI: 10.1007/s11192-014-1280-0

How to evaluate the degree of interdisciplinarity of an institution?

Lorenzo Cassi · Wilfriedo Mescheba · Élisabeth de Turckheim

Abstract The Stirling index of the set of references of the corpus documents is widely used in the literature on interdisciplinary research and is defined as the integration score of the corpus under study. Such an indicator is relevant at the scale of a research institution, however, there is a gap between the integration scores of individual documents, and a global score computed on the whole set of references. The difference between the global index and the average of individual document indexes carries another relevant information about the corpus: it measures the diversity between the reference profiles of the corpus documents. It is, therefore, named *between article* index whereas the average of the individual article indexes is called *within article* index. The statistical properties of these two indexes as well as of the global index are derived from a general approximation method for distributions and lead to statistical tests which can be used to make meaningful comparisons between an institution indexes and benchmark values. The two dimensions of the global index provide a more acute information on the interdisciplinary practices of an institution researchers in a given research domain and is, therefore, likely to contribute to strategic and management issues.

Keywords Interdisciplinary research · Indicators · Stirling index decomposition · Inertia of a set of weighted points · Asymptotic distribution of indicators · Statistical tests

Mathematics Subject Classification (2000) MSC 60F05

L. Cassi · W. Mescheba · E. de Turckheim
OST - Observatoire des Sciences et des Techniques, 75015 Paris, France
E-mail: (lorenzo.cassi, wilfriedo.mescheba, elisabeth.deturckheim)@obs-ost.fr

L. Cassi
Université Paris 1, Centre d'économie de la Sorbonne, 75013, Paris, France
Address for correspondence: lorenzo.cassi@univ-paris1.fr

É. de Turckheim
INRA, Délégation à l'évaluation, 75007 Paris, France
E-mail: elisabeth.deturckheim@paris.inra.fr

Introduction

In a seminal article, **?** perform an empirical exercise which aims at evaluating the interdisciplinarity of a researchers set of articles in terms of integration degree. To do so, they wonder: "how should we compute I[ntegration score] for him or her?" (p.136). According to them, the "obvious choice" is to take the average of the integration score of each article co-authored by the researcher. Is this choice so obvious when an institution research activity has to be evaluated?

In a more recent article, **?** seem to think that it is not. In their comparison among three economics departments and three innovation studies units of UK universities in terms of interdisciplinarity of their research output, they calculate an overall integration score for each institution based on the whole set of references instead of an average of the integration scores of the institutions publications. The difference is far from being trivial, because in the former case the interdisciplinarity between publications is taken into account while in the latter the interdisciplinarity depends exclusively on the interdisciplinarity of each publication.

Which choice is more relevant? What are we actually measuring (or pretending to measure) in one case rather than the other? The aim of our paper is to answer these questions. Indeed, we mathematically analyse the relation between the global and the average indexes. Moreover, we develop a set of statistical tests in order to evaluate if the difference between an observed value and a benchmark value is significant.

The rest of the paper is organized as follows: section "Measuring interdisciplinarity" reviews the different measures of interdisciplinarity used in the literature and justifies the choice of Stirling index[1]. Section "Global index *deconstruction*" presents the decomposition of the global index and section "Choosing weights for the references" proposes two options for the weighting of the references of each document. In section "Giving a meaning to index values: statistics for the Stirling indexes" we propose two probabilistic models to associate the Stirling indexes to random variables and compute their distributions. This allows to design statistical tests to compare the indexes for an institution to reference values. Finally, we have two empirical sections. In the first one, "Impact of the weight choice: some empirical evidence", we analyse the differences in terms of results of the weighting method and try to get some conclusion about the best choice in the context of interdisciplinary studies. In the second empirical section "A case study: a French University" we perform an empirical analysis for a research institution. By doing so, we show the insights our approach can provide for strategic analysis and for research institutions policy follow-up. In the "Discussion" we report a few issues brought up by this work related to the application of this framework for science policy studies and also about methodological choices. In appendix A, we recall the

[1] "Stirling index has become known in ecology literature as Rao's quadratic entropy" (**?**, p.267 footnote 4).

definition of the inertia of a set of points and show how its decomposition can be translated to the Stirling index. While in appendix B, we provide the proofs for the convergence of the empirical indexes and their use for statistical inference.

Measuring interdisciplinarity

Following the literature on the topic, we adopted the definition of interdisciplinarity proposed by National Academies in a 2005 report (**?**): a mode of research that integrates different disciplines in order "to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single field" (p. 119). Interdisciplinarity is therefore defined in terms of cognitive content of the research activity.

In order to capture information on research interdisciplinarity one needs, therefore, to consider the characteristics of research output such as scientific articles and their references. In this context, we limit our review to bibliometric approach, without considering other methods such as peer-review or content analysis. We acknowledge that bibliometric quantitative measures are not able to grasp the complexity of the phenomenon and that "they leave considerable gaps in understanding" as emphasized recently by **?** in their review of the state-of-the-art. However, these methods allow systematic comparison, large scale analysis and they are relatively less time-consuming than, for instance, direct inquiries based on primary data. That explains also why analysing directly the output of research activity is still one of the most diffused methodologies[2].

In the following, we briefly review the main measures of interdisciplinarity presently available and discussed in the literature. However, independently from their differences, all of them deal with the concept of diversity. The more an article integrates sources from different disciplines, the more it is interdisciplinary.

Indeed it is not possible to conceive interdisciplinarity without first defining disciplines, i.e. a disciplinary taxonomy. This is true with both a top-down (i.e. predefined categories) and a bottom-up approach (i.e. articles clustered into factors based on some similarity measure analysis). Moreover, the degree of interdisciplinarity depends on "how finely or coarsely discipline categories are defined" (**?**, page 5). The narrower discipline boundaries are, the greater interdisciplinary is observed since a piece of research would more likely cross disciplines in terms of sources, methods and tools. In the literature, the majority of analyses refers to scientific categories provided by Web of Science, Thomson Reuters. However, from the analytical point of view, it is enough to have a partition of all articles (or journals) into blocks that can correspond to disciplines, fields or categories. So, we refer generally to term category for the sake of simplicity but without a specific taxonomy in mind.

[2] One of the main conclusions of the already cited survey of **?** is that this kind of analysis should be completed by other methods in order to fill the gaps in understanding.

Given a corpus of publication, and in order to the diversity of categories, three different types of data can be dealt with. First, we can focus on the publications themselves and their distribution among categories. Secondly, it is possible to examine article references, i.e. the citations given by an article. Finally, we can take into account the categories of articles citing the article under study. As interdisciplinarity is understood as an integration process of different sources, literature tends to use references as the main informative data. Diversity of references can actually grasp how the scientific process has combined previous results and "logically seems the best gauge of intellectual integration" (**?**, page 127). So in the following we will consider the diversity of the set of article references in order to measure the article interdisciplinarity degree.

As claimed among other by **?**, diversity contains three different dimensions: (1) variety: number of distinctive categories; (2) balance: evenness of distribution; and finally (3) disparity: degree to which the categories are different.

In order to achieve a quantitative approach, the choice of a metrics is the next issue to address. All the measures proposed in the literature can be consequently classified according to which of the three dimensions they are able to capture and how. An example of measure capturing only the first dimension, variety, is the number of distinctive categories to which belong the references of an article. Two classical measures, such as Shannon's index[3] and Simpson's index[4], are able to combine variety and balance dimensions but fail to account for disparity between different categories (**?**).

As pointed by Stirling in his 2007 paper, starting from a flexible general heuristic is a more relevant approach than seeking a definitive diversity index[5]. He shows that the heuristic D

$$D = \sum_{i,j=1}^{m} (d_{ij})^{\alpha} (p_i p_j)^{\beta}$$

complies with eight desirable requirements for a diversity index. Moreover, D also allows different weightings on balance, variety and disparity through the choices for α and β [6].

Following the same choice as most recent papers on interdisciplinarity studies[7], we use the integration score derived from the heuristic D with $\alpha = 1$ and

[3] $SH = \sum_i p_i \log p_i$, where p_i is the proportion of elements in category i

[4] $SI = 1 - \sum_i pi^2 = \sum_{i \neq j} p_i p_j$.

[5] We thank A. Stirling for drawing our attention to this point in a mail exchange on a previous version of the paper.

[6] In the formula for D, p_i is the proportion of elements in category i and d_{ij} is the distance between categories i and j, m is the number of categories and α and β are two parameters to choose between 0 and 1.

[7] An exception is the choice of Vincent Larivière and Yves Gingras, in (**?**) and in other work, who use the percentage of references in other categories than the citing article as an interdisciplinarity measure.

$\beta = 1$. The index we use -and call Stirling index- is thus

$$ST = \sum_{i,j=1}^{m} p_i p_j d_{ij}.$$

In this index, the joint contribution of two categories i and j i.e. $p_i p_j$, unlike other indexes (e.g. Simpson), is weighted with d_{ij}. This weight captures the degree to which the categories are different and allows taking into account the epistemic distance between two categories. Another interesting implication of this choice is that if two categories have a distance equal to zero, the sum of their separate contributions to the value of the diversity index would be the same as the contribution of a single category in which the two categories would be merged. In that sense, Stirling index can correct the categories definition. This means that taking into account disparity (i.e. distance between categories) makes the Stirling index robust relatively to the adopted taxonomy. Therefore, Stirling index, unlike other indexes, overcomes at least partially the issue of an arbitrary choice of a predefined categorisation.

Global index *deconstruction*

The assessment of the interdisciplinarity of a corpus can be performed at two different levels: (1) calculating the diversity index for the references of each single article and taking their average value or (2), at the more aggregated level, calculating a global diversity index of the concatenated set of the references of the articles belonging to the corpus. Of course a global index is relevant if the corpus is reasonably homogeneous as for instance a set of publications in a given domain or discipline where the practice of citation is similar. This corpus, denoted $\mathcal{A}$ contains n documents hereafter called articles. These articles cite N documents of a set $\mathcal{R}$, containing possibles ties when some references are common to more than one article of $\mathcal{A}$. The global index is the Stirling index of this set $\mathcal{R}$.

If we consider a single article a and its set of references $\mathcal{R}_a$, the Stirling index of this set is the interdisciplinarity index of article a, denoted ST_a. If we denote N_a the number of references of article a and N_{ai} the number of those references which are in category i, the Stirling index for the article a is then

$$ST_a = \sum_{i,j} \frac{N_{ai}}{N_a} \frac{N_{aj}}{N_a} d_{ij}\,.$$

The global index ST is neither the sum nor the average of article indexes, because there is a component of the global diversity of $\mathcal{R}$ which is not captured in the individual indexes as underlined by **?**. The missing component consists in the diversity of references between articles. Using the similarity of the Stirling index with the inertia of the cloud of points associated to the documents in $\mathcal{R}$ as explained in the appendix "Inertia of a set of points", the global diversity

index can be decomposed into two terms corresponding to the diversity of the reference *within* each article and the diversity *between* the sets of references of the different articles as

$$ST = ST^W + ST^B.$$

Given this result, it is possible to analyse separately the two components of the global index. The *within* component counts for the diversity of the knowledge base of each article; the *between* component measures how much the articles belonging to the same corpus are diverse from each other in terms of knowledge sources. For instance, the set of publications of an institution can be evaluated as highly interdisciplinary either because researchers publish highly interdisciplinary articles with very similar disciplinary profile of their references or because their articles are highly specialized, i.e. with references in a small number of close disciplines, but very much different from each other in terms of their references. These two polar scenarios can in principle display the same global measure of interdisciplinarity. Our approach, different from the previous ones, is able to distinguish such cases and bring further insights into the assessment of the interdisciplinary of a corpus of publications.

In order to interpret the value of an index of an institution, a benchmark is needed. In a study where the interdisciplinarity of the different research fields of an institution is the issue at stake, a benchmark for each field could be the world scientific production in the same field. As in **?**, the value of the interdisciplinarity indexes of the world production in the same field and the same period of time will be considered as reference values. The difference between an index computed on the corpus of the institution publications and the corresponding index for the reference corpus provides a relevant indicator. We show in section "Giving a meaning to index values: statistics for the Stirling indexes" how to decide whether a difference is significant or not and, for the time being, we propose a qualitative interpretation of these indicators, which fully exploits the decomposition of the global index into its two components.

For a given corpus of publications and for each of the two components, two situations are possible: the corpus of the institution publications in a domain or a discipline has an index smaller or larger than the index of the reference corpus. A smaller *within* index means that the institution publications are more specialized - regarding the range of disciplines of their references - than the average of the publications in the reference corpus[8]. On the contrary, a larger *within* index means that the institution publications are more integrative. Concerning the *between* index, a smaller index for the institution means that the research is focused around specific niches while a larger *between* index corresponds to a larger range of diversity of the institution research in the domain considered. Combining the cases for the two components leads to define a taxonomy with four types summarised in boldface characters in Table 1. However, as we explain it further, errors should be taken into account and a

[8] Here, the meaning of *specialization* is not the same as in **?** where specialization means a low diversity of the categories of the journals where the articles are published. In this paper, the word is used to characterize the low diversity of the references of the articles.

Table 1 Taxonomy for an institution research domains based on the values of its *within* and *between* indexes

Between index of a research domain	**Within** index of a research domain		
	smaller than benchmark	*equivalent to benchmark*	*larger than benchmark*
larger than benchmark	**Wide variety of research with specialized articles**	Wide variety of research with standard article specialization	**Wide variety of research with integrative articles**
equivalent to benchmark	Standard variety of research with specialized articles	Standard variety of research with standard article specialization	Standard variety of research with integrative articles
smaller than benchmark	**Niche research with specialized articles**	Niche research with standard article specialization	**Niche research with integrative articles**

small difference between institution and reference corpus indexes would not be meaningful. Therefore a median line and a median column should be included in Table 1 and this leads to a taxonomy with nine cases.

The position of each research domain of an institution in this taxonomy is likely to shed some light on the inclination towards interdisciplinary in the different domains and therefore to provide clues for strategic thinking policy and management.

Choosing weights for the references

Decomposing the global index into its two components requires making a choice for the weight of each reference and hence for the weight of each article. Two choices are possible: (1) equal weight to each reference irrespective of the number of references of the article which cites it, (2) equal weight to the set of references of each article. If we assign a same weight to each reference of the whole set $\mathcal{R}$ of the corpus under study, articles with large number of references such as review articles, will have a larger contribution to the global index than articles with fewer references as, for instance, articles dealing with specific issues[9]. In this case, the within component of the global index decomposition will be a weighted sum of the article indexes

$$ST^W = \frac{1}{N}\sum_{a=1}^{n} N_a ST_a \,.$$

[9] This is apparently the implicit assumption made by **?** when they calculate the global index of diversity for the corpuses of different institutions.

We call **EWR** this weight option where all references have the same weight, irrespective of the number of references of the citing article.

A second option assigns a same global weight equal to 1 to the set $\mathcal{R}_a$ of references of each article a, which means that a reference weight is the inverse of the number N_a of references of the citing article.
We denote **EWA** this weight option where each article equally contributes to the global index as well as in both the within and between components. In this case, the within index is simply the average of articles indexes

$$ST^W = \frac{1}{n}\sum_{a=1}^{n} ST_a$$

and the proportions q_i to define the global index

$$ST = \sum_{i,j=1}^{m} q_i q_j d_{ij}$$

are not computed from the categories proportions of the set $\mathcal{R}$ but with the averages of the proportions of the sets $\mathcal{R}_a$

$$q_i = \frac{1}{n}\sum_{a=1}^{n} \frac{N_{ai}}{N_a}.$$

For either weight choice, it is easy to calculate the global and the within diversity indexes, and subtracting the latter from the former to obtain the between article index as well. We show in the appendix that the between component corresponds to the diversity (or inertia) of n points, each one representing a kind of average category (or centre of gravity) of the references of an article.

Giving a meaning to index values: statistics for the Stirling indexes

When the issue is to describe science and to compare how the different disciplines collaborate, it is natural to say that a discipline is more interdisciplinary than another one when the Stirling index for the set the publications in this discipline displays a greater value than the other discipline (**?**). Similarly, an institution could be considered as more interdisciplinary than another institution in the same domain if its global index value is greater than the global index of the other institution (**?**). Is such a conclusion drawn from a simple observation of the differences satisfactory? The issue is to control the variability of each of the three indexes, for example the variation due to bibliometric errors as wrong assignment of articles to categories etc. To cope with this issue, **?** use a Student's t-test to compare the degree of interdisciplinarity of a discipline over time. However other authors generally do not report statistical significance of the comparisons.

In the present analytical framework where six indexes are proposed, only the within index in the EWA case can be dealt with the standard Student's t-test because it is a simple average of the article indexes. The five other cases require additional statistical arguments to define tests equivalent to the Student's t-test.

To fill this methodological gap, we first define a two-step probability model for each weight option:

- In the EWR case, the first random step is to select an article a and the value N_a of the number of its references. The second step selects the N_a categories of the references so that each category i is drawn with probability p_i.
- In the EWA case, the first step selects both the number of references N_a and the parameters $p_a = (p_{a1}, \ldots, p_{ai}, \ldots, p_{am})$. The probability distribution of p_a is such that the mean value of p_{ai} is q_i. In the second step the N_a categories of the references are selected so that each category i is drawn with probability p_{ai}.

In the appendix, we compute the asymptotic distribution of the three indexes (global, within and between) which are derived from an extended version of the central limit theorem for functions of averages of independent random variables. Estimating the variance of the limit distribution allows to define normalised statistics (i.e. with unitary variance) associated with each of the six indexes. Statistical tests to compare the index values for the institution to reference values are therefore available. They rely on the normal asymptotic distribution of the difference between the observed (random) index and the (deterministic) reference value, divided by the variance of the observed index. Practically, as we show in section "A case study: a French University", the value of the normalized test statistic (or z-score) provides a scale that can be used for graphical representations of the *within* and *between* components of interdisciplinarity of an institution in the different research disciplines where the institution is active. Comparisons between two institutions are also possible with the two-sample versions of the same statistical tests.

Impact of the weight choice: some empirical evidence

Two options for the weight were given to each reference and hence to each article. Choosing which of the two is more correct may prove to be a tricky issue. In this section we propose to look for an empirical answer because it is not obvious to provide a priori arguments in favour of one option rather than the other.

Intuitively, we are inclined to think that differences induced by weight choice are mainly related with the distribution of the number of references per article. It is plausible that the index ST_a of an article increases with the number of references. For instance, articles with very few references (e.g. 1 or 2) are likely to have a low index. Thus, with that hypothesis, the more unbalanced the

distribution of the number of references is, the larger the expected difference between the two weighting options will be.

Indeed, an a priori argument follows: the EWA option seems less arbitrary, because the number of references of a paper depends on the authors' practice that is in principle not related with their interdisciplinary orientation. Consequently, the EWA weighting option may be preferred because it is not affected by individual citation behaviour, e.g. the number of references.

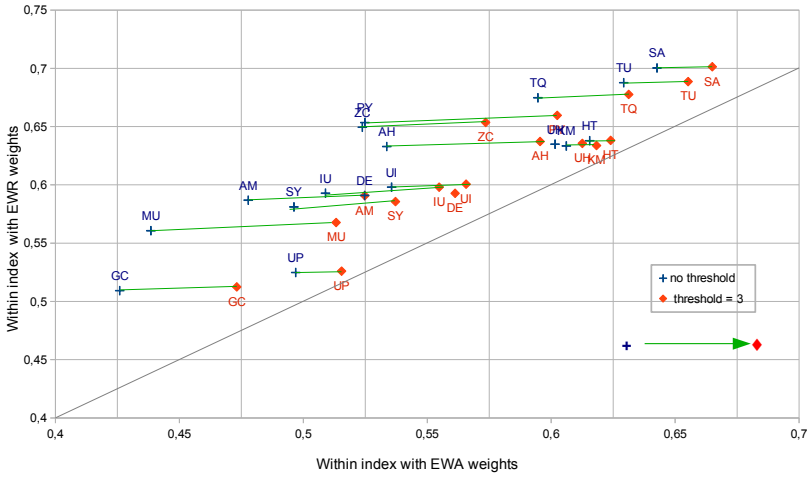

Fig. 1 *Within* indexes for the two weighting options with two thresholds: no threshold (blue crosses), threshold of 3 references (red diamonds)

In order to provide some evidence, we calculated each index for the two weighting options for a sample of 80 scientific categories for all WoS documents (article, letter, review and note) published in 2008. In the two figures displayed, for the sake of readability, we selected a representative subset of the 80 categories which index values cover the whole range of observed values. Moreover, for each index, between or within, we displayed two values for each category: for the first one any publication with at least one reference in WoS was included (no threshold) while for the second only the publications with at least three references in WoS were taken into account (threshold = 3)[10].

Two main empirical results are : (1) the difference between the two weighting options, (2) the effect of a threshold.

[10] We understand that **?** who base the index calculation on articles "containing at least three citations to journals articles that link to a subject category" (page 278) use the same threshold condition. In another paper on the same topic, the calculation includes the documents "with at least three cited SCs that appear in WoS" (**?**, page 137) which is a different condition, somewhat weaker since a single reference is often associated with more that one WoS category.

Figure 1 shows that for the within index the value for EWR is greater than for EWA for all selected categories [11]. This can be explained if articles with a few references have a smaller value of ST_a as the average will be lower in EWA case where these articles have a higher weight than in EWR case. As shown in Figure 2, we have the opposite result for the between index as the value for EWA is greater than for EWR for all selected categories. The explanation is much less intuitive than for ST^W. It could be that articles with few references have particular reference profiles, more distant from the average article profile. As these articles weigh more in the EWA case, the average distance between article profiles (which is the between index) will be larger than in the EWR case. However, this possible explanation deserves further exploration.

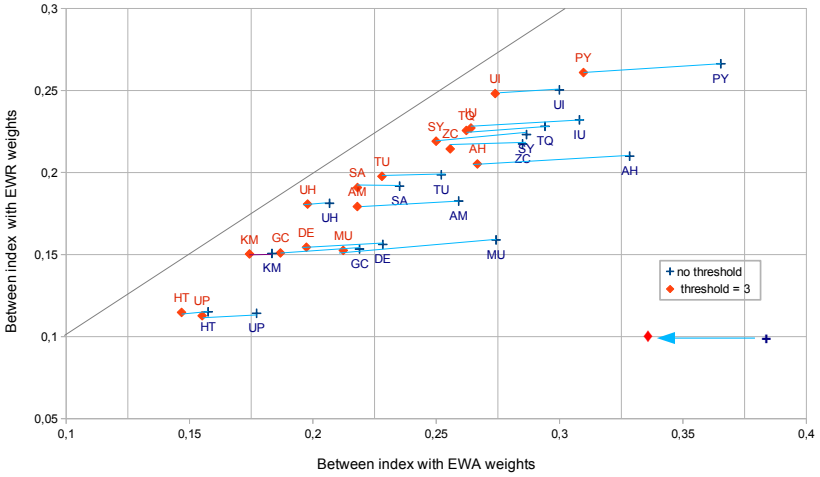

Fig. 2 *Between* indexes for the two weighting options with two thresholds: no threshold (blue crosses), threshold of 3 references (red diamonds)

Let's us consider now the effect of the threshold. Both figures show that setting a threshold of a least three references lead to less divergent values between the two weighting options. This essentially happens because the within index increases for EWA and the between index decreases whereas the two indexes are fairly constant for the EWR case, as shown by the horizontal shift of the points. We suspect that raising further the value of the threshold (e.g. at least four references) would make the difference between the two weights even smaller, but a higher threshold should probably be adapted to each discipline. However, the two indexes would not be completely equal. A fraction of the difference would probably be caused by articles with a very large number of references, as review articles. Therefore, setting a threshold deals with the issue

[11] This is true for any of the 80 selected categories.

of the left distribution tail, but the issue of the right tail of the distribution is not solved by choosing a minimum value for the number of references. Mainly for this reason, we think that EWA option selecting articles with a minimum number of references is a more appropriate solution than the EWR choice because it copes with both issues of articles with either very small or very large number of references which would bias the values of interdisciplinarity indexes.

A case study: a French University

The aim of this section is to show how the scores of the Stirling indexes can be used to build an interdisciplinarity map of the institution research domains. In a preliminary step of an ongoing study with a group of a dozen of French universities, a domain has been defined as the research published in journals of a WoS category. The 15 categories with the largest publication counts have been selected for each university and the interdisciplinarity map of the categories has been provided. The preliminary results for one of these universities are displayed below.

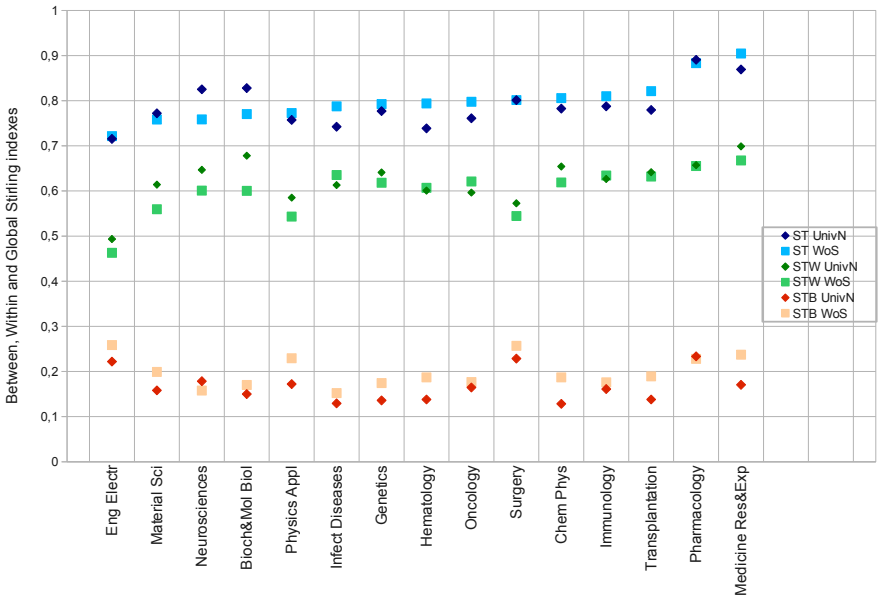

Fig. 3 Values of the *global, within* and *between* indexes (series appearing respectively in top, middle and bottom position) for the 15 selected WoS categories. WoS categories are ordered by increasing values of the WoS global index *ST*.

Data for this case study are the university publications of the four types as follows: article, review, letter and note that were published between 2008 and 2012. They have been extracted by OST from Thompson Reuters Web of

Science (WoS) database which is available at OST and then validated by the university. Only the publications with at least three references in the WoS have been included. The benchmark values are the indexes for all WoS publications of the same types of publications of the chosen categories published during 2008 and with at least three references in WoS. The weight choice is EWA. Data treatments were achieved at OST with SAS software.

As shown in Figure 3, the index values are strongly dependent on the category and the differences between the institution and the benchmark are smaller than the variation between categories: this confirms the need for choosing a relevant benchmark.

For the categories of this study, the global index decomposition is about 3/4 into the *within* component and 1/4 into the *between* component. This means that there is a rather large homogeneity of the references profiles inside one category. Conversely, a case where the averages profiles (centres of gravity) of references of the articles would be more diverse than their diversity by article leads to suspect the relevance of the set of journals to represent a discipline.

The scores for the global index show that the 15 categories are arranged in 3 groups (Fig. 4): a group of 7 categories with negative scores (Group 1), a group of categories with absolute values of the indexes lower than 2.58 and therefore not significantly different at level 1% from the WoS (Group 2) and a group of 2 categories with extremely high scores (Group 3).

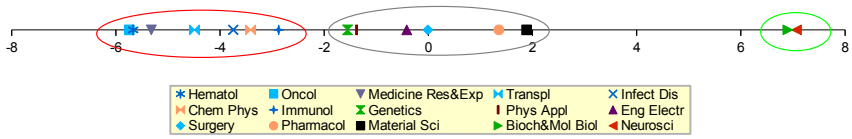

Fig. 4 Scores of the global indexes for the comparison of the Stirling indexes of the university under study with the WoS indexes for 15 WoS categories.

The decomposition of the global index provides some insight on the characterisation of the interdisciplinarity of the university research in each WoS category. Figure 5 displays the scores of the within and between index scores on the horizontal and vertical axes. Shaded zones defined by parallel lines to the axes correspond to non significant differences at level 1% between the university and the reference indexes.

The three groups of Figure 4 resulting from the global score of are delimited in Figure 5 by ellipses ranging from bottom left to top right. Pairs of disciplines as for instance *Material Sciences* and *Pharmacology* in Group 2 which have very similar global indexes correspond to very different combinations of the *within* and *between* components: the research in pharmacology of this university is fairly similar to the world standard while the research in material sciences in this university is probably positioned in a few niches

corresponding to the scientific strategy of the research teams. Moreover, these teams probably have a multidisciplinary network of collaborations. The position of such categories in the right low part of the graph may indicate a comparative advantage of the university in some topics of these disciplines.

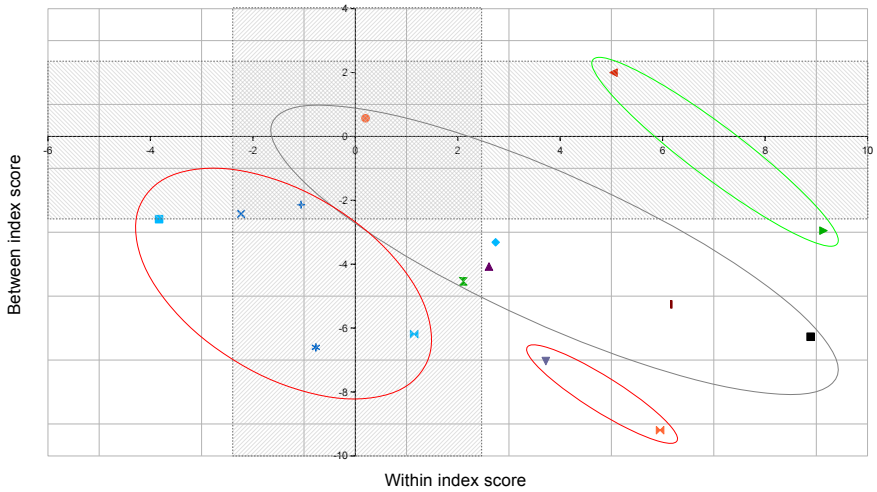

Fig. 5 Scores of the *within* and *between* indexes for the comparison of the Stirling indexes of the university under study with the WoS indexes. Groups of Fig 4 are shown in ellipses with the corresponding colour. Icons for categories are the same as in Fig 4. Shaded zones correspond to scores which are not significant at level 1%.

Finally, we can observe that there is no category with significant indexes in the upper part on the graph, which would correspond to a value of the *between* index greater for the institution than for the WoS. This is a more general remark than this particular case: such values are rarely observed and this is consistent with the fact that an institution would likely have a smaller diversity between its publications than the whole diversity existing at the world scale.

Discussion

In this paper we define a framework that allows analysing the degree of interdisciplinarity of a research institution. To do so, we focus on two important dimensions characterising the research profile of an institution: the average diversity of the references of the articles (i.e. ST^W) and the diversity among articles reference profiles (i.e. ST^B). Computing these indexes and comparing them to a benchmark in a systematic way provides meaningful insights on research activity. However, the use of the Stirling indexes as a tool for the evaluation of institutions strategies is an issue which has to be further explored.

First, these indexes can be used as a positioning tool for the different research domains of an institution within the international research landscape. Beyond the comparison with the world standard, comparisons between different institutions are likely to provide interesting information and to raise research policy issues. Second, the effect of social and organisational factors on interdisciplinarity can be analysed as well. For example, comparing institutions working on the same topics with different internal organisations or providing different type of support for social interaction between teams and departments could provide insights on the impact of different strategies.

For such comparisons, a relevant corpus of publications of the institution under study has to be defined. For instance, in the case study of the previous section, we define the corpus of a university in terms of WoS categories but this is not a refined choice. Categories may be too large because they could blend journals with very different scopes into the same corpus. In that sense, a relative importance of the *between* component could signal a problem in the definition of categories as a too wide category. An alternative would be to use an improved classification or, in particular studies, to refer social/institutional boundaries as a department production instead of a corpus determined by a category. For instance, in **?**, they consider the field of Innovation studies and Business management in their comparison of different institutions. Indeed, the choice of the corpus on which the indexes are computed is to be carefully considered in each study. There is no methodological difficulty to do so as soon as the corresponding benchmark can be determined, which is the case if the corpuses are defined in terms of journals.

Another issue, in principle independent from the previous one, is the choice of the WoS categories as the classification frame to compute the Stirling index. As claimed by **?**, global maps of science are largely insensitive to the basic classification into categories and, provided the classification has the same level, maps look very much similar. The same should be true for the Stirling index. However, possible doubts could be removed in splitting categories into more homogeneous ones with respect to their reference profile. Obviously the categories could as well be journals as **?** and **?** did to produce relevant classifications and maps of science. However, as the Stirling index is insensitive to the merge of categories with equal reference profiles, such a refined classification may not be strictly necessary.

Another remark is related to the choice of the Stirling index which may be considered as arbitrary. In our framework, we use this index for comparisons and not as absolute values. If only changes of the index are considered as variations over time, across similar institutions or across different research domains, the choice of a particular index is less stringent and other indexes would be appropriated as well. However, an argument for choosing Stirling index is that this index can be decomposed into its *within* and *between* components much easier than another index. This relies on the presence of a distance in the index definition and more precisely because the distance d_{ij} is chosen as the cosine between the vectors representing categories i and j and is therefore the square of an Euclidean distance (section A.1).

Acknowledgements The authors are grateful to Ismael Rafols and Tommaso Ciarli for a fruitful discussion on an earlier version of this paper. The issues presented in the section "Discussion" largely result of this interaction. In particular, I. Rafols made us aware of the relevance of analysing interdisciplinarity at the level of a department rather than at the university level. This issue will be considered in future work with different institutions.

Appendices

A Decomposition of the Stirling index

In this section, we show that, when the distance d_{ij} between categories i and j used in the Stirling index is the square of an Euclidean distance, it is possible to interpret the Stirling index as the inertia of a set of points associated to the documents of a corpus (theorem 1). When this corpus can be split into subsets, the decomposition of the inertia into an inertia *within* the subsets and an inertia *between* the subsets provides a decomposition of the Stirling index of interdisciplinary into two components (theorem 2).

A.1 Inertia of a set of points

In mechanics, the dynamics of a solid depends on the weights of the elementary components of the solid and on their distances to the centre of gravity. We use the following definition for a the solid which is a set of weighted points:

Definition 1 The inertia I of a set $\mathcal{R} = \{A_1, \ldots, A_N\}$ of N points with weights $w_1, \ldots, w_N$ is

$$I = \sum_{r,s=1}^{N} w_r w_s \, \delta^2(A_r, A_s) \tag{1}$$

where $\delta(A_r, A_s)$ is the distance between A_r and A_s.

The two following properties will be useful.

Lemma 1 *If G is the centre of gravity of a set $\mathcal{R}$ of N weighted points, the inertia of $\mathcal{R}$ is such that*

$$I = 2W \sum_{r=1}^{N} w_r \, \delta^2(A_r, G) \tag{2}$$

where $W = \sum_r w_r$ is the total weight of $\mathcal{R}$.

Proof If the components of the vector $\mathbf{A}_r$ are the coordinates of the point A_r in the Euclidean space with the δ metrics, and if the associated norm of $\mathbf{A}_r$ is denoted $\|\mathbf{A}_r\|$ and the scalar product of the two vectors $\mathbf{A}$ and $\mathbf{B}$ is denoted $\mathbf{A} \cdot \mathbf{B}$,

$$\begin{aligned} \delta^2(A_r, A_s) &= \|\mathbf{A}_r - \mathbf{A}_s\|^2 \\ &= \|\mathbf{A}_r - \mathbf{G}\|^2 + \|\mathbf{G} - \mathbf{A}_s\|^2 + 2(\mathbf{A}_r - \mathbf{G}) \cdot (\mathbf{G} - \mathbf{A}_s) \end{aligned}$$

As G is the centre of gravity of points A_r, $\sum_r w_r(\mathbf{A}_r - \mathbf{G}) = 0$. The sum over the two indexes r and s gives the expression of I as in (2).

Lemma 2 *If the set $\mathcal{R}$ is split into n subsets $\mathcal{R}_a$ of respectively N_a points, the inertia of $\mathcal{R}$ is the sum of the average of the inertia I_a of the subsets $\mathcal{R}_a$ and the inertia $\widetilde{I}$ of the set $\widetilde{\mathcal{R}}$ of the n centres of gravity of the subsets $\mathcal{R}_a$*

$$I = \sum_{a=1}^{n} \frac{W}{W_a} I_a + \widetilde{I} \tag{3}$$
$$\widetilde{I} = \sum_{a,b=1}^{n} W_a W_b \, \delta^2(G_a, G_b)$$

where $W_a = \sum_{A_r \in \mathcal{R}_a} w_r$ is the weight of $\mathcal{R}_a$ and G_a is the centre of gravity of $\mathcal{R}_a$.

Proof Splitting the sum in formula (2) into n terms associated with the n subsets $\mathcal{R}_a$,

$$\begin{aligned} I &= 2W \sum_{a=1}^{n} \sum_{A_r \in \mathcal{R}_a} w_r \delta^2(A_r, G) \\ &= 2W \sum_{a=1}^{n} \sum_{A_r \in \mathcal{R}_a} w_r \left(\delta^2(A_r, G_a) + \delta^2(G_a, G) \right) \end{aligned}$$

and recognizing I_a, written as in (2), in each term of the sum over a leads to

$$I = 2W \sum_{a=1}^{n} \frac{1}{2W_a} I_a + 2W \sum_{a=1}^{n} W_a \, \delta^2(G_a, G).$$

Because the centre of gravity of $\mathcal{R}$ is also the centre of gravity of the set $\bar{\mathcal{R}}$ of the centres of gravity of the subsets $\mathcal{R}_a$

$$\widetilde{I} = 2W \sum_{a=1}^{n} W_a \delta^2(G_a, G)$$

is the inertia of the set $\widetilde{\mathcal{R}}$ of the points $G_1, \ldots, G_a, \ldots, G_n$ with weights $W_1, \ldots, W_a, \ldots, W_n$ where $W_a = \sum_{A_r \in \mathcal{R}} w_r$.

A.2 Decomposing the Stirling index as the inertia of a set $\mathcal{R}$

The set $\mathcal{R}$ is now a set of points representing the N documents referenced by a corpus $\mathcal{A}$ of n documents (hereafter called articles). $\mathcal{R}$ is the sum of the sets $\mathcal{R}_a$ and each $\mathcal{R}_a$ is associated with the N_a references of article a. Each reference is represented by a point A_i which only depends on the discipline i of the reference. The distance between A_i and A_j is $\delta(A_i, A_j)$.

The two Stirling indexes considered in this paper are associated with two different choices of the weights w_r:

1. **EWR choice** with equal weights for references : $w_r = 1$ and therefore W and W_a are the numbers N and N_a of points in $\mathcal{R}$ and in $\mathcal{R}_a$,
2. **EWA choice** with equal weights for articles : for a point A_r associated with a reference of article a, $w_r = N_a^{-1}$ so that $W_a = 1$ and $W = n$.

In the rest of this document we use subscripts for the notations of the three Stirling indexes when the formulas are different in the two cases. Thus ST_R, ST_R^W , ST_R^B stand for the global,

within and between Stirling indexes in the EWR case, ST_A, ST_A^W, ST_A^B for the same indexes in the EWA case. For instance, we denote the global indexes presented on page 5 and 8

$$ST_R = \sum_{i,j} p_i p_j d_{ij}$$
$$ST_A = \sum_{i,j} q_i q_j d_{ij}.$$

In the **EWR case**, if there are N_i^+ references in discipline i, the inertia of the set $\mathcal{R}$ is

$$I_R = \sum_{i,j=1}^{m} N_i^+ N_j^+ \, \delta^2(A_i, A_j).$$

Moreover, if the distance d_{ij} used in the Stirling index is such that $d_{ij} = \delta^2(A_i, A_j)$, we can write the inertia as

$$I_R = \sum_{i,j=1}^{m} N_i^+ N_j^+ \, d_{ij} = N^2 \sum_{i,j=1}^{m} p_i p_j d_{ij} = N^2 ST_R.$$

In the **EWA case**, the location of the points A_r representing references only depend on the discipline of the reference, and their weights only depend on the article a which cites them. There are N_{ai} references in discipline i cited in article a. In formula (1), we group the terms A_r and A_s corresponding to references in the same disciplines i and j cited by the same articles a and b. Therefore,

$$I_A = \sum_{a,b,i,j} N_{ai} N_{bj} w_a w_b \, \delta^2(A_i, A_j).$$

As $w_a = N_a^{-1}$, we have

$$\begin{aligned} I_A &= \sum_{a,b,i,j} \frac{N_{ai}}{N_a} \frac{N_{bi}}{N_b} \, \delta^2(A_i, A_j) \\ &= \sum_{i,j} \delta^2(A_i, A_j) \sum_a \frac{N_{ai}}{N_a} \sum_b \frac{N_{bi}}{N_b} \\ &= n^2 \sum_{i,j} q_i q_j \, \delta^2(A_i, A_j) \\ &= n^2 ST_A \end{aligned}$$

A general formula for I, which also holds for $\widetilde{I}$, is then

$$I = W^2 ST. \tag{4}$$

We summarize this relation between the Stirling index and an inertia in the following theorem.

Theorem 1 *The Stirling index ST of a set $\mathcal{R}$ of documents based on a classification of $\mathcal{R}$ into m classes is equivalent to the inertia I of the set - also denoted $\mathcal{R}$ - of weighted points A_r representing these documents in an Euclidian space, where documents from the same class are represented with points with the same location. Provided the distance δ in the Euclidian space and the distance d of the Stirling index are such that $d_{ij} = \delta_{ij}^2$, the inertia of $\mathcal{R}$ is*

$$I = W^2 ST$$

where W is the sum of the weights of the points in $\mathcal{R}$.

We note that, when restricted to $\mathcal{R}_a$, this formula is written $I_a = W_a^2 ST_a$ which is, for either weight choice, consistent with the definition of ST_a as in page 5

$$ST_a = \sum_{i,j} \frac{N_{ai}}{N_a} \frac{N_{aj}}{N_a} d_{ij} \,. \tag{5}$$

Applying the decomposition of the inertia of formula (3) and replacing I with $W^2 ST$, we obtain a decomposition for the Stirling index in the two choices of weights:

Theorem 2 *The global Stirling index ST of the set $\mathcal{R}$ of all references of a corpus $\mathcal{A}$ under study, can be split into two terms: an index ST^W measuring the disciplinary diversity within the references of each article and an index ST^B measuring the diversity of the references between the articles of the corpus:*

$$\begin{aligned} ST &= ST^W + ST^B \\ ST^W &= \frac{1}{W} \sum_{a=1}^{n} W_a ST_a \\ ST^B &= \frac{1}{W^2} \sum_{a,b} W_a W_b \, \delta^2(G_a, G_b) \end{aligned}$$

The 'within index' is a weighted average of the Stirling indexes ST_a associated with the n individual articles and the 'between index' is the Stirling index of a set $\widetilde{\mathcal{R}}$ of n points $G_1, \ldots, G_a, \ldots, G_n$, each of them representing the virtual 'average' discipline of the references of an article.

We note that, in the EWA case, ST^W is a simple average of the article indexes

$$ST_A^W = \frac{1}{n} \sum_{a=1}^{n} ST_a$$

whereas in the EWR case, the *within* Stirling index ST^W is a weighted average of article indexes

$$ST_R^W = \frac{1}{N} \sum_{a=1}^{n} N_a ST_a.$$

The *between* index is just computed by difference as $ST^B = ST - ST^W$.

A.3 Euclidean representation of $\mathcal{R}$

If we use the same distance between disciplines as **?** and **?**, this distance d_{ij} is derived from a similarity index s_{ij} which is the cosine of two vectors $\mathbf{A}_i$ and $\mathbf{A}_j$. If the citing direction is considered to define the distance between two categories, the vectors $\mathbf{A}_i$ and $\mathbf{A}_j$ are derived from columns of the matrix $N^* = (N^*_{k,i})$ where $N^*_{k,i}$ is the number of references in discipline k cited by the publications in discipline i of the reference corpus. If the cited direction is used, $\mathbf{A}_i$ and $\mathbf{A}_j$ are lines of the matrix N^*. If $\mathbf{A}_i$ is normalized so that its norm is such that $\|\mathbf{A}_i\|^2 = \frac{1}{2}$, the distance d_{ij} between categories is the square of the Euclidean distance between the end points A_i and A_j of the vectors $\mathbf{A}_i$ and $\mathbf{A}_j$ as follows

$$\begin{aligned} \delta^2(A_i, A_j) &= \|\mathbf{A}_i - \mathbf{A}_j\|^2 \\ &= \|\mathbf{A}_i\|^2 + \|\mathbf{A}_j\|^2 - 2\cos(\mathbf{A}_i, \mathbf{A}_j)\|\mathbf{A_i}\|\|\mathbf{A_j}\| \\ &= 1 - \cos(\mathbf{A}_i, \mathbf{A}_j) = d_{ij}. \end{aligned}$$

Therefore, an Euclidean space exists where d_{ij} is the square distance between points A_i and A_j representing the categories i and j so that the inertia of the set $\mathcal{R} = \{A_1, \ldots, A_N\}$ is the Stirling index associated with the proportions of elements of $\mathcal{R}$ in the categories.

B Statistical properties of the Stirling indexes

B.1 Notations for the statistical section

For this section dealing with statistical issues, we need to use different notations for the empirical and the theoretical (or true) values of the different indexes. The formulas of page 5 and 8 for the global index ST will now have two versions. The true values will be denoted ST_R^0 and ST_A^0

$$ST_R^0 = \sum_{i,j} p_i p_j d_{ij}$$
$$ST_A^0 = \sum_{i,j} q_i q_j d_{ij},$$

and their empirical versions denoted

$$\widehat{ST_R} = \sum_{i,j=1}^{m} \hat{p_i}\hat{p_j} d_{ij} \tag{6}$$
$$\widehat{ST_A} = \sum_{i,j=1}^{m} \hat{q_i}\hat{q_j} d_{ij} \tag{7}$$

where

$$\hat{p_i} = \frac{1}{N} \sum_a N_{ai}$$
$$\hat{q_i} = \frac{1}{n} \sum_a \frac{N_{ai}}{N_a}.$$

Under the probability models defined page 9, the theoretical (or true) values of the global index can be related with empirical versions of the index.
In the EWR case, the two-step probability model is such that the first random selection is to draw an article a or equivalently to draw the value of N_a. The second step selects N_a elements of $\mathcal{C}$ such that the m dimensional variables $(N_{a1}, \ldots, N_{am}), a = 1, \ldots, n$ are n independent multinomial variables of parameters N_a and $p = (p_1, \ldots, p_m)$, where the parameter p is common to the n variables. The conditional mean of N_{ai} given N_a is $p_i N_a$. Therefore the (unconditional) means are

$$E(N_{ai}) = p_i E(N_a)$$
$$E(\hat{p_i}) = p_i.$$

In the EWA case, the two-step probability model is such that the first random choice selects an article a or equivalently values for N_a and for
$p_a = (p_{a1}, \ldots, p_{ai}, \ldots, p_{am})$. The distribution of p_a is such that

$$E(p_a) = (q_1, \ldots, q_m).$$

The articles are independently selected which means that the $(m+1)$ dimensional variables $(N_a, p_a), a = 1, \ldots, n$ are independent and so are the variables $N_{ai}/N_a, a = 1, \ldots, n$. The second step selects N_a elements of $\mathcal{C}$ with a multinomial distribution of parameters N_a and p_a. The conditional distribution of $(N_{a1}, \cdots, N_{ai}, \cdots, N_{am})$ given N_a and p_a is a multinomial distribution with parameters N_a and p_a. Therefore the conditional mean of N_{ai}/N_a is p_{ia} and its (unconditional) mean is q_i

$$E(\hat{q_i}) = q_i.$$

This relates the empirical and theoretical values of the global index.

Concerning the within index, the formulas on page 19 correspond to empirical values and are now denoted $\widehat{ST_A^W}$ and $\widehat{ST_R^W}$

$$\widehat{ST_A^W} = \frac{1}{n}\sum_{a=1}^{n} ST_a \tag{8}$$

$$\widehat{ST_R^W} = \frac{1}{N}\sum_{a=1}^{n} N_a ST_a. \tag{9}$$

They are related to the corresponding theoretical values as follows:

$$\begin{aligned} ST_A^{W0} &= E(ST_a) \\ ST_R^{W0} &= \frac{E(N_a ST_a)}{E(N_a)}. \end{aligned}$$

Finally, for the between index, the theoretical values are, like the empirical values, computed as the differences

$$\begin{aligned} ST_A^{B0} &= ST_A^0 - E(ST_a) \\ ST_R^{B0} &= ST_R^0 - \frac{E(N_a ST_a)}{E(N_a)}. \end{aligned}$$

B.2 A general central limit theorem

For statistical inference based on the empirical value of an index, we need to compute the asymptotic distribution of variables such as

$$\sqrt{n}\left(\widehat{ST} - ST^0\right).$$

This is achieved for large samples when a central limit theorem ensures that this distribution is asymptotically normal with mean zero and a variance σ^2 which can be estimated. Then a normalized statistic or z-score denoted z^*

$$z^* = \frac{\sqrt{n}}{\hat{\sigma}}\left(\widehat{ST} - ST^*\right)$$

can be used to compare the true value ST^0 to a reference value ST^*.

The expected central limit theorem are derived from a general theorem for a smooth function f of averages of independent, identically distributed random variables. This theorem is based on a the delta-method which uses the first order Taylor expansion of the function f. This theorem is available with different formulations in various teaching documents as the lecture notes in Statistics by **?**, theorem 8.9, page 221[12]. We reformulate the general theorem with convenient notations for our situation.

Theorem 3 (Delta method) *Let $(X_{1a}, X_{2a}, \ldots, X_{ma})$, $a = 1, 2, \ldots, n$, denote a n-sample from a m-dimensional distribution and consider a function $f = f(X_1, X_2, \ldots, X_m)$ of the averages*

$$X_i = \frac{1}{n}\sum_{a=1}^{n} X_{ia},$$

[12] Our first use of the method was the formulation proposed by **?** for the empirical covariance which is very close to the probabilistic result needed for ST_A^B.

then, for f sufficiently smooth, the variables

$$W_n = \sqrt{n}\,(f(X_1, X_2, \ldots, X_m) - f(E(X_1), E(X_2), \ldots, E(X_m))$$
$$W_n^1 = \sqrt{n}\sum_{i=1}^{m} \lambda_i (X_i - E(X_i))$$

have the same asymptotic distribution which is normal with mean zero and variance σ_F^2, where the coefficient λ_i is the partial derivative of f at $(E(X_1), E(X_2), \ldots, E(X_m))$

$$\lambda_i = \frac{\partial f}{\partial x_i}(E(X_1), E(X_2), \ldots, E(X_m))$$

and σ_F^2 is the variance of $F_a = \sum_i \lambda_i X_{ia}$.

The theorem follows from the fact that, when the second order derivatives of f are bounded, the function $f(X_1, \ldots, X_n) - f\,(E(X_1), E(X_2), \ldots, E(X_m))$ and its first order expansion $\sum_i \lambda_i (X_i - E(X_i))$ have the same asymptotic distribution.

B.3 Asymptotic distribution of the global index

In the EWA case, if we denote $\hat{q}_i = X_i$, according to (7), $\widehat{ST_A}$ is a function of the averages X_i as

$$\widehat{ST_A} = h((X_i)_{i=1,\ldots,m}) \quad \text{where} \quad h((x_i)_{i=1,\ldots,m}) = \sum_{i,j} d_{ij} x_i x_j$$

and

$$\sqrt{n}\left(\widehat{ST_A} - ST_A^0\right) = \sqrt{n}\,(h((X_i)_i) - h((E(X_i)_i))\,.$$

We use the partial derivatives of h

$$\lambda_i = 2\sum_j d_{ij} q_j = 2\gamma_i$$

where

$$\gamma_i = \sum_j d_{ij} q_j$$

to apply theorem 3 to get the asymptotic distribution of $\sqrt{n}\left(\widehat{ST_A} - ST_A^0\right)$.

Corollary 1 *The distribution of the global index in the EWA case is such that the asymptotic distribution of*

$$\sqrt{n}\left(\widehat{ST_A} - ST_A^0\right) = \sqrt{n}\sum_{i,j=1}^{m} d_{ij}\,(\hat{q}_i\hat{q}_j - q_i q_j)$$

is a centered normal distribution with a variance σ_H^2 which is the variance of

$$H_a = 2\sum_i \gamma_i \frac{N_{ai}}{N_a} \tag{10}$$

where $\gamma_i = \sum_j d_{ij} q_j$.

For statistical purposes, we estimate γ_i with $\widehat{\gamma_i} = \sum_j d_{ij}\hat{q_j}$ and σ_H^2 with the empirical variance $\widehat{\sigma_H^2}$ of

$$\widehat{H_a} = \sum_i 2\widehat{\gamma_i}\frac{N_{ai}}{N_a}.$$

In the EWR case, according to (6), the variable $\widehat{ST_R}$ is a function of the averages

$$Y_i = \frac{1}{n}\sum_a N_{ai}$$

$$Y = \frac{1}{n}\sum_a N_a$$

$$\widehat{ST_R} = k((Y_i)_i; Y) \quad \text{where} \quad k((y_i)_i; y) = \sum_{i,j} d_{ij}\frac{y_i y_j}{y^2}.$$

As $E(Y_i) = p_i E(N_a)$, we have

$$k((E(Y_i))_i; E(Y)) = \sum_{i,j} d_{ij}p_i p_j$$

so that

$$\sqrt{n}\left(k((Y_i)_i; Y) - (k(E(Y_i))_i; E(Y))\right) = \sqrt{n}\left(\widehat{ST_W} - ST_W^0\right).$$

We use the partial derivatives of k

$$\lambda_i = \frac{2}{E(N_a)}\sum_j d_{ij}p_j = \frac{2}{E(N_a)}\beta_i$$

$$\lambda = -\frac{2}{E(N_a)}\sum_{i,j} d_{ij}p_i = -\frac{2SR_R^0}{E(N_a)}$$

with

$$\beta_i = \sum_j d_{ij}p_j$$

to apply theorem 3 to get the asymptotic distribution of $\sqrt{n}\left(\widehat{ST_R} - ST_R^0\right)$.

Corollary 2 *The distribution of the global index in the EWR case is such that the asymptotic distribution of*

$$\sqrt{n}\left(\widehat{ST_R} - ST_R^0\right) = \sqrt{n}\sum_{i,j=1}^{m} d_{ij}\left(\hat{p_i}\hat{p_j} - p_i p_j\right)$$

is a centered normal distribution with variance σ_K^2 which is the variance of

$$K_a = \frac{2}{E(N_a)}\left(\sum_i \beta_i N_{ai} - ST_R^0 N_a\right) \tag{11}$$

and where $\beta_i = \sum_j d_{ij}p_j$.

For statistical purposes, we estimate β_i with $\widehat{\beta_i} = \sum_j d_{ij}\hat{p_j}$, $E(N_a)$ with Nn^{-1}, ST_R^0 with $\widehat{ST_R^0}$ and σ_K^2 with the empirical variance $\widehat{\sigma_K^2}$ of

$$\widehat{K_a} = \frac{2n}{N}\left(\sum_i \widehat{\beta_i}N_{ai} - \widehat{ST_R}N_a\right).$$

B.4 Asymptotic distribution of the within index ST^W

In the EWA case, according to (8), $\widehat{ST_A^W}$ is the average of the n independent variables ST_a, a standard central limit theorem for $n^{-1}\sum_a ST_a$ ensures that

$$\sqrt{n}\left(\widehat{ST_A^W} - ST_A^{W0}\right) = \sqrt{n}\left(\frac{1}{n}\sum_a ST_a - E(ST_a)\right)$$

converges to a normal centered variable with variance equal to $\mathrm{Var}(ST_a)$. Therefore the usual Student's test is applicable to compare the mean (or theoretical value) of the empirical $\widehat{ST_A^W}$ to a reference value ST_A^{W*} or to compare the theoretical values associated with two independent samples.

Corollary 3 *The distribution of the within index in the EWA case is such that the asymptotic distribution of* $\sqrt{n}\left(\widehat{ST_A^W} - ST_A^{W0}\right)$ *is normal with mean zero and the same variance as* ST_a.

In the EWR case, according to (9), $\widehat{ST_R^W}$ is a function of the averages S and Y defined as follows

$$S = \frac{1}{n}\sum_a N_a ST_a$$
$$Y = \frac{1}{n}\sum_a N_a$$

$$\widehat{ST_R^W} = q(S,Y) \quad \text{where} \quad q(s,y) = sy^{-1}.$$

As the value of q at the variable mean is the theoretical value of $\widehat{ST_R^W}$,

$$q(E(N_a ST_a), E(N_a)) = \frac{E(N_a ST_a)}{E(N_a)} = ST_R^{W0}$$

and the partial derivatives of q are

$$\lambda_1 = \frac{1}{E(N_a)}$$
$$\lambda_2 = -\frac{E(N_a ST_a)}{E(N_a)^2} = -\frac{ST_R^{W0}}{E(N_a)}$$

we just apply theorem 3 to get the asymptotic distribution of $\sqrt{n}\left(\widehat{ST_R^W} - ST_R^{W0}\right)$.

Corollary 4 *The distribution of the within index in the EWR case is such that the asymptotic distribution of* $\sqrt{n}\left(\widehat{ST_R^W} - ST_R^{W0}\right)$ *is normal with mean zero and a variance* σ_Q^2 *which is the variance of*

$$Q_a = \frac{1}{E(N_a)}\left(N_a ST_a - ST_R^{W0} N_a\right) \tag{12}$$

For statistical purposes, we estimate $E(N_a)$ with Nn^{-1}, ST^{W0} with $\widehat{ST_R^W}$ and σ_Q^2 with $\widehat{\sigma_Q^2}$ the empirical covariance of $\widehat{Q_a} = nN^{-1}\left(N_a ST_a - \widehat{ST_R^W} N_a\right)$.

B.5 Asymptotic distribution of the between index ST^B

As $ST^B = ST - ST^W$, this variable is written as a function of averages of independent variables which are, with the same notations as above, $((X_i)i, \frac{1}{n}\sum_a ST_a)$ in the EWA case, and $((Y_i)i, Y, S)$ in the EWR case. The first order Taylor expansion of $\sqrt{n}\left(\widehat{ST^B} - ST^{B0}\right)$ is just the difference of the first order expansions of $\sqrt{n}\left(\widehat{ST} - ST^0\right)$ and $\sqrt{n}\left(\widehat{ST^W} - ST^{W0}\right)$. Therefore the asymptotic distribution of $\sqrt{n}\left(\widehat{ST^B} - ST^{B0}\right)$ is straightforward and we get the two asymptotic distributions.

Corollary 5 *The distribution of the between index in the EWA case is such that the asymptotic distribution of*

$$\sqrt{n}\left(\widehat{ST_A^B} - ST_A^{B0}\right)$$

is normal with mean zero and variance σ_U^2, where σ_U^2 is the variance of U_a

$$U_a = H_a - ST_a$$

where H_a is as in (10).

Corollary 6 *The distribution of the between index in the EWR case is such that the asymptotic distribution of*

$$\sqrt{n}\left(\widehat{ST_R^B} - ST_R^{B0}\right)$$

is normal with mean zero and variance σ_V^2, where σ_V^2 is the variance of U_a

$$V_a = K_a - Q_a$$

where K_a is as in (11) and Q_a as in (12).

To get the test statistics, we replace H_a, K_a and Q_a with their estimates.

B.6 Asymptotic correlation of ST^W and ST^B

In order to measure the probabilistic dependence of the three indexes, it is possible to calculate their asymptotic correlations. This is also a corollary of theorem 3. For instance, the asymptotic covariance of ST^W and ST^B in the EWA case is the covariance of ST_a and U_a

$$\begin{aligned}\lim_n \operatorname{cov}(ST_A^W, ST_A^B) &= \operatorname{cov}(ST_a, U_a)\\ \lim_n \operatorname{cov}(ST_R^W, ST_R^B) &= \operatorname{cov}(Q_a, V_a).\end{aligned}$$

Unfortunately, the probabilistic models defined above do not lead to simple values of these covariances. In general, there is no probabilistic independence between the two indexes. However, there might be particular distributions of (N_a, ST_a) or of (K_a, Q_a) where the two indexes are uncorrelated. To test whether this is the case for a given corpus, a normalised statistic can be derived with the same delta method.

B.7 Asymptotic variance of the category contributions to the global index

For a refined analysis, the global index ST can be splitted into contributions of individual categories according to $ST^0 = \sum_i C_i^0$, which is, in the EWA case for instance

$$C_i^0 = q_i \sum_j q_j d_{ij} \, .$$

To evaluate the asymptotic variance of

$$\widehat{C_i} = \hat{q}_i \sum_j \hat{q}_j d_{ij}$$

we use theorem 3 and we write $\widehat{C_i} = h^{(i)}((X_i)_{i=1,\dots,m})$ where $X_i = \hat{q}_i$ and

$$h^{(i)}((x_i)_{i=1,\dots m}) = x_i \sum_j x_j d_{ij} \, .$$

As the partial derivatives of $h^{(i)}$ are

$$\begin{aligned} \lambda_j^{(i)} &= x_i d_{ij} \quad \text{if} \quad j \neq i \\ \lambda_i^{(i)} &= \sum_j x_j d_{ij} = \gamma_i \end{aligned}$$

the asymptotic variance of $\sqrt{n}\left(\widehat{C_i} - C_i^0\right)$ is the same as the variance of

$$F_a^{(i)} = \sum_j q_i d_{ij} \frac{N_{aj}}{N_a} + \gamma_i \frac{N_{ai}}{N_a} \, .$$

As usual, we estimate the variance of $F_a^{(i)}$ with the empirical variance of $\widehat{F_a^{(i)}}$ where

$$\widehat{F_a^{(i)}} = \mathbf{D^{(i)}} \cdot \mathbf{P_a} \tag{13}$$

where

$$\begin{aligned} D_j^{(i)} &= \hat{q}_i d_{ij} \quad \text{if} \quad i \neq j \\ D_i^{(i)} &= \hat{\gamma}_i \\ P_{aj} &= \frac{N_{aj}}{N_a} \, . \end{aligned}$$